Handling Variation in Speech and Language Processing

Handling variation in speech and language processing (request change to: Handling variation in spoken language processing)

Simon King

Revised: December 2, 2004

Abstract

When spoken language systems are faced with variations such as different speakers or the presence of noise, their performance usually degrades. This chapter looks at some of the ways such systems handle variation and how they attempt to prevent this degradation in performance.

1 INTRODUCTION

The companion article *Language variation in speech technologies* describes the many sources and types of variation in spoken language that affect speech technology systems. In this article, we look at how such systems handle this variation.

1.1 Why is variation generally seen as a problem?

Most speech and language processing systems involve some sort of pattern matching, whether that is explicit in terms of rules or finite state machines, or implicit via a model, such as the hidden Markov model. These rules or models are usually tailored to some data set (for statistical models, this is called training). Variation poses two distinct problems in this case.

Firstly, variation in the training data means that the rules or model must have sufficiently broad coverage to fit all of the data, and this can lead to a system in which distinctions between classes become less clear. In the case of modelling data using statistical distributions, this means larger variances, which in turn leads to more overlap between the distributions associated with different classes. This will lead to poorer performance in classification of unseen ("test") data.

Secondly, if the test data varies with respect to the training data, those rules or models which may be a good fit to the training data will not be such a good fit to the test data; again, this leads to poorer

performance in terms of accuracy.

So, it appears that variation causes serious problems for these systems: more variation will almost always lead to worse performance. How can systems deal with variation, and in particular, how can accuracy be maintained even in the face of variation?

1.2 Further reading

These articles are suggested reading on speech technology: *Language variation in speech technologies*; *Speech Technologies, Overview*; *Natural Language Processing, Overview*; *Speech Recognition, Statistical Methods*. The textbooks listed in the bibliography offer further reading and references back to the original research papers for each technique: (Jurafsky & Martin, 2000) is recommended as a general introduction; (Holmes & Holmes, 2001) has more detail; (Huang, Acero & Hon, 2001) is the most comprehensive. Suggested reading on variation includes: *Variation and Language, an overview*; *Corpora studies of variation*; *Accent*; *Dialect Atlases*; *Dialect surveys*; *Subjective and objective measures of variation.*

2 SPEECH RECOGNITION

There are three distinct approaches to handling variation in automatic speech recognition (ASR) systems. It can be essentially ignored, the system can try to remove it through some pre-processing step, or the statistical model can be adapted to better model the noisy data. Often, many approaches are applied in a single system (Furui, 2001: 343).

2.1 Ignoring variation

Since the ubiquitous model for ASR, the hidden Markov model, is a statistical model (*see* Speech Recognition, Statistical Methods), variation in training data does not prevent use of the usual training methods, such as Baum-Welch parameter estimation.

The most common form of variation is corruption of the speech signal with additive noise or convolutional (channel) noise. A simple approach to building a recogniser that will work with noisy data is to train it on similar noisy data. For example, a recogniser for telephone-quality speech would be best trained on such speech, and not on clean speech collected in a recording studio. However, any variation in the noise between training and test data will lead to a significant degradation in accuracy; in situations where the type of noise varies (e.g. recognition of mobile telephone speech), such an approach will not be successful. Because noisy speech data is difficult to collect, it is often artificially

constructed by adding noise to clean speech, and/or passing the speech down a real or simulated telephone line.

When the type of noise is not known in advance, the system may be trained on a data set consisting of sub-sets, each of which contains a different type of noise. The resulting system will be more robust to previously unseen noise types.

Of course, there are other types of variation than additive or convolutional noise. Variations within and between speakers, between differing microphones, and in the type of speech (e.g. read text versus spontaneous conversation) also present challenges. It is possible to ignore all of these and simply train the system on data which matches the test data as closely as possible. However, when this is impractical or impossible, such as when the variation changes over time, other techniques must be employed

2.2 Removing variation from the signal

For speech transmission systems, like telephone networks, noise must be removed from the speech signal to improve intelligibility for the listener. However, in ASR it is not necessary to produce a clean speech waveform, only a parameterisation of the signal, such as a sequence of cepstral vectors (*see* Speech Recognition, Statistical Methods). Therefore, techniques can be applied during or after parameterisation of the waveform.

Cepstral mean subtraction After parameterisation, normalisation is commonly applied. Even simple forms of normalisation such as cepstral mean normalisation, in which the long-term average of the cepstral vectors is subtracted from the cepstral vector for each frame, can be highly effective for removing channel variation and some speaker variation (Furui, 2001: 341).

Vocal tract length normalisation (VTLN) A difference in vocal tract length, such as between males and females, causes a shift in the formant space. A simple way of compensating for this is to apply a speaker-specific frequency warping during parameterisation. The warping factor can be estimated automatically from a small sample of speech.

Echo cancelling and dereverberation In telephony, signals are reflected back from the receiver along the telephone line, causing echos which impair intelligibility. Automatic echo cancellation is employed to remove these signals. Similar techniques may be applied before ASR. Echos also occur when speech waves are reflected off surfaces prior to being recorded by the microphone. The delay

time between direct and reflected versions of the speech arriving at the microphone is much shorter than the echoes in telephony. This *reverberation* causes serious problems for ASR systems.

RASTA filtering During parameterisation, filtering can be applied. In RASTA filtering, bandpass filters are applied to remove both very slow and very rapid variations in the parameters, because these variations cannot be attributed to the speech signal (Hermansky & Morgan, 1994).

2.3 Adapting hidden Markov models

It is common for a system to have a large, well-trained set of hidden Markov models (HMMs) but be faced with test data which does not match the models well enough to enable high-accuracy recognition. There may be either speaker or noise mismatch. One approach taken in this type of situation is to *adapt* the models to the test data. This requires some adaptation data, for which we may or may not know the correct word sequence. To use methods which require a known word sequence in situations where this information is not available, an initial recognition pass can be used to give a noisy (errorful) estimate of the word sequence.

Maximum a posteriori (MAP) Given some data with a known word sequence, the HMM parameters can be adjusted to maximise the likelihood of the new parameter values, given the adaptation data. This method can only adapt the parameters of the HMMs involved in modelling the particular word sequence of the adaptation data – which will generally be a very small subset of all the system's parameters. Therefore, this method tends to be very slow.

Maximum likelihood linear regression (MLLR) With only a limited amount of data, it is not feasible to retrain all parameters of a large set of HMMs. MAP adaptation only adjusts the parameters of those models corresponding to the adaptation data; in contrast, MLLR applies a transformation to all the parameters of the entire system at once. This transform is linear and there may be either a single global transform matrix or a number of transforms applied to different classes of models. The transform can be learned from only small amounts of adaptation data.

Parallel model combination This technique is only useful for adapting to noise, rather than speaker variation. If a model is available for the noise alone, it may be combined with a model for clean speech to produce a model of noisy speech. To obtain the model of noise at recognition time, it is necessary

to locate parts of the signal without speech and use them to train the model. When the speech level relative to the noise (signal-to-noise ratio, SNR) is low, this is not trivial to do automatically.

2.4 Adapting the language model

Language models for ASR are usually N-gram models (*see* Language Processing, Statistical Methods) trained on very large amounts of data (e.g. 100 million words). Since there is no speech corpus sufficiently large, these models are invariably trained on text corpora, which usually means either newspaper text or text drawn from the world wide web. This text data will differ from transcriptions of speech, so some form of adaptation may be applied to make the models a better match for spoken language.

Pooling training data One simple technique is to pool large amounts of text data with small amounts of speech transcriptions. When computing the N-gram counts, the speech data is usually more heavily weighted to compensate for the small quantity available.

Language model interpolation An alternative to pooling data is to train two language models, one on each type of data, then interpolate the probabilities predicted by each model. The interpolation weights can be set to maximise the probability of some *held-out* data not used during training.

Cache-based adaptation Within a single passage of speech, certain words will be temporarily much more likely than was observed in the training data. Cache-based models exploit this: words seen so far (in the recogniser output) are stored in a cache and used to train a low-order language model (a bigram or unigram). The probabilities predicted by this cache model are then interpolated with those from a general model.

Topic adaptation Adaptation can also help in situations where the speech to be recognised changes *topic* from time to time, such as in broadcast news, causing sudden changes in the probabilities of some words. For example, as the topic shifts from sports results to the stock market, the probabilities of football players' names will decrease and the probabilities of company names will increase. To adapt to a given topic, we need some more data from this topic to train the language model. Information retrieval (*see* Document Retrieval, Automatic) techniques can be used to find this data.

2.5 Preventing variation at the source

It is sometimes simpler and more effective to prevent variation at the source, rather than attempt to remove it later.

Choice of microphone Close-talking microphones produce a much better signal-to-noise ratio than lapel or desk-mounted ones. Directional or noise cancelling microphones are designed to be much more sensitive to signals coming from a single direction (the speaker) than to ones coming from other directions (background noise) (Huang, Acero & Hon, 2001: section 10.2).

Microphone arrays In some situations, the use of multiple microphones can help reduce the amount of noise captured by the system (Huang, Acero & Hon, 2001: section 10.4). Microphone arrays consist of two or more microphones spaced some distance apart. Just as humans use binaural hearing to localise sounds, a microphone array uses the arrival time difference of a signal between the various microphones to localise sounds. By *beamforming*, the array can accept signals from just one direction, and reject signals from other directions. Microphone arrays can also help remove reverberation by beamforming towards only the direct-path speech signal.

Dialogue design In practical dialogue systems deployed today, the machine always has the initiative (*see* Dialogue). This allows the system designers to manipulate the dialogue in order to dramatically reduce the difficulty of the ASR task. For example, whilst a human telephone banking teller may ask the customer "How may I help you?", an automated system may ask "What sort of transaction would you like to make?". By effectively limiting the range of possible answers, the system is able to impose a much more restrictive language model during ASR, and hence achieve a lower error rate. The language model will be switched at each turn in the dialogue.

3 SPEECH SYNTHESIS

System architecture By carefully structuring the architecture of a text-to-speech (TTS) (*see* Speech Synthesis) system, it is possible to partition the linguistic resources required by the various modules such that maximal re-use can be achieved. For example: the statistical models and/or rules used for expanding abbreviations may be shared across all accents of a language; the phoneme inventory may be shared across all speakers of a particular accent.

Pronunciation dictionary Writing pronunciation dictionaries for tens of thousands of words by hand is extremely time-consuming and therefore expensive. It also takes a great deal of expert knowledge and, because consistency is important, cannot usually be distributed to a large team of people. However, a close match between the speaker and the pronunciation dictionary is crucial. To resolve this conflict, a method has been developed to generate accent- or speaker-specific dictionaries from a single language-specific but accent-independent "meta-dictionary" using a relatively small set of rules and settings (Fitt & Isard, 1999).

Pitch marking Most concatenative speech synthesis systems (*see* Speech Synthesis) require accurate pitch marking of the speech waveforms in the inventory. It is difficult to construct an automatic pitch marking algorithm that works "out of the box" for all speakers. Usually, some parameters, like the minimum and maximum expected pitch period durations, must be tuned to the individual speaker.

4 OTHER SPEECH TECHNOLOGIES

Whilst variation is almost always a problem for ASR, sometimes variation between speakers is actually very useful.

Speaker identification Identifying who is talking is a useful task in its own right (*see* Speaker Recognition and Verification, Automatic). The obvious application is security, e.g. telephone banking, building access control. Two distinct tasks can be identified: speaker identification ("Which one of a set of speakers is talking?") and speaker verification ("Is the speaker who they claim to be?"). Somewhat surprisingly, the parameterisation of speech used for speaker-independent ASR is often the most effective one for speaker identification or verification, and similar statistical models are used (HMMs). However, discriminative training techniques (*see* Speech Recognition, Statistical Methods) are usually required to get high accuracy.

Speaker tracking In multi-party speech, such as in a business meeting, we may wish to simultaneously recognise the speech of all speakers, and annotate the resulting output with speaker identity. Since speakers overlap a great deal, we must use the different characteristics of each individual speaker (e.g. F0) to do this (possibly in combination with speaker location information obtained from a microphone array).

Speaker characteristic identification Rather than the actual identity of the speaker, we may only be interested in certain characteristics. For example, a system may try to guess the gender of a speaker, and then use gender-specific models for recognition. Since most ASR systems are language-specific, a language guesser must be employed as an initial step when multiple languages are to be recognised. Automatically determining the accent of a speaker has applications in dialectometry (*see* Dialect Atlases).

5 Conclusion

Whilst variation, from whatever source, can cause major problems for spoken language systems, a wide range of techniques are available to minimise the degradation in performance. These include: preventing variation; removing it from speech signals; transforming parameterised speech; adjusting the parameters of the statistical models.

Bibliography

Fitt, S. & Isard, S. (1999) *Synthesis of regional English using a keyword lexicon*. In Proc. Eurospeech '99, volume 2, pages 823-826, Budapest.

Furui, S. (2001). *Digital Speech Processing, Synthesis and Recognition*, 2nd edition. Marcel Dekker.

Gold, B. & Morgan, N. (1999). *Speech and Audio Signal Processing*. Wiley.

Hermansky, H. & Morgan, N. (1990). *RASTA Processing of Speech*, IEEE Transactions on Speech and Audio Processing, **2**(4), pages 578-589.

Holmes, J. N. & Holmes, W. J. (2001). *Speech Synthesis and Recognition*, 2nd edition. Taylor & Francis

Huang, X., Acero, A., Hon, H-W. (2001). *Spoken language processing: a guide to theory, algorithm and system development*. Prentice Hall.

Jelinek, F. (1998). *Statistical Methods for Speech Recognition*. MIT Press.

Jurafsky, D. & Martin, J. (2000). *Speech and language processing*. Prentice Hall.